Computerlinguistik und Massenüberwachung

State-of-the-Art nach Snowden-Fundus, INDECT-Papers & Co.

Hernani Marques

Chaos Singularity 2015 @ Biel/Bienne

13. Juni 2015

Was ist Computerlinguistik?

Was ist “Computerlinguistik” für ***dich***?



Definition gem. Universität Zürich UZH[1]

Die Computerlinguistik ist eine junge Disziplin im Überschneidungsbereich von Sprachforschung und Informatik. Sie untersucht,

- *wie die menschliche Sprache als Mittel zur Übermittlung, Speicherung und Verarbeitung von Information verwendet wird, und*
- *wie man diese Prozesse auf dem Computer modellieren und für konkrete Anwendungen nutzbar machen kann.*

Dies geschieht primär aus theoretischem Interesse. [...]

[1] Vgl. Webseite: `http://www.cl.uzh.ch/what-is-cl.html`

Definition auf der Wikipedia DE[2]

In der Computerlinguistik (CL) oder linguistischen Datenverarbeitung (LDV) wird untersucht, wie natürliche Sprache in Form von Text- oder Sprachdaten mit Hilfe des Computers algorithmisch verarbeitet werden kann. Sie ist Schnittstelle zwischen Sprachwissenschaft und Informatik.

[2] Vgl. Webseite: `https://de.wikipedia.org/w/index.php?title=Computerlinguistik&oldid=139308216`

Definition auf der Wikipedia EN[3]

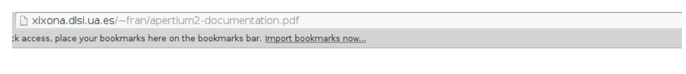
Computational linguistics is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective.
[...]
Computational linguistics has theoretical and applied components, where theoretical computational linguistics takes up issues in theoretical linguistics and cognitive science, and applied computational linguistics focuses on the practical outcome of modeling human language use.

[3] Vgl. Webseite: `https://en.wikipedia.org/w/index.php?title=Computational_linguistics&oldid=663599748`

- Tokenisierung
- Wortartenerkennung
- Named Entity Recognition (NER)
- Parsing (by Anerkennung einer Syntax in einem Text)
- Koreferenzauflösung
- Kollokationen

- Sentiment Analysis
- Machine Translation (regelbasiert, statistisch oder hybrid)
- Text Mining / Relationship Mining
- Automatic Text Summarization
- Authorship Recognition
- Topic Analysis

Beispiel Pipeline Apertium (Maschinelle Übersetzung)



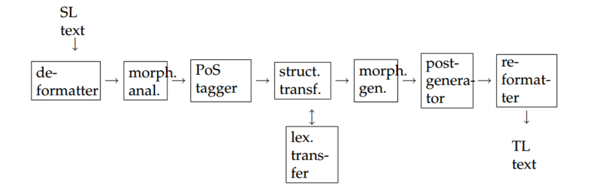
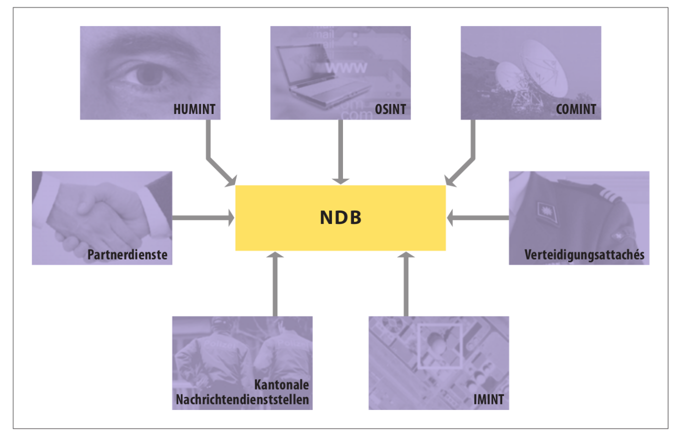


Figure 1.1: The eight modules that build the assembly line of the shallow-transfer machine translation system.

Die Sensoren des NDB



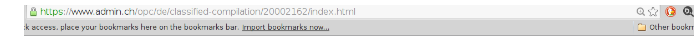
Onyx-Funkaufklärung



Onyx-Auswertung



6-Monate-VDS via EJPD



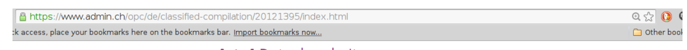
Art. 12 Pflichten der Anbieterinnen

[1] Die Anbieterinnen von Postdiensten sind verpflichtet, der anordnenden Behörde die Postsendungen sowie die weiteren Verkehrs- und Rechnungsdaten soweit herauszugeben, als es in der Überwachungsanordnung umschrieben wird. Sie erteilen der anordnenden Behörde auf Verlangen weitere Auskunft über den Postverkehr einer Person.

[2] Sie sind verpflichtet, die Daten, welche eine Teilnehmeridentifikation erlauben, sowie die Verkehrs- und Rechnungsdaten während mindestens sechs Monaten aufzubewahren.

[3] Die Tatsache der Überwachung und alle sie betreffenden Informationen unterliegen gegenüber Dritten dem Post- und Fernmeldegeheimnis (Art. 321ter StGB[1]).

5-Jahres-VDS via VBS



Art. 4 Datenbearbeitung

[1] Das ZEO vernichtet die im Rahmen der Funkaufklärung gewonnenen Resultate spätestens im Zeitpunkt der Beendigung des jeweiligen Funkaufklärungsauftrags.

[2] Es vernichtet die erfassten Kommunikationen spätestens 18 Monate nach deren Erfassung.

[3] Es vernichtet die erfassten Verbindungsdaten spätestens 5 Jahre nach deren Erfassung.

[4] Es darf Daten, die aufgrund eines Funkaufklärungsauftrags erfasst worden sind, auch zur Erfüllung eines anderen Funkaufklärungsauftrags des gleichen Auftraggebers verwenden.

[5] Die Anmeldung von Datensammlungen, das Auskunfts- und Einsichtsrecht sowie die Archivierung richten sich nach den für den jeweiligen Auftraggeber geltenden rechtlichen Bestimmungen.

Beispiel: Massenüberwachung Onyx (2)

www.parlament.ch/d/dokumentation/berichte/berichte-delegationen/berichte-der-geschaeftspruefungsdelegation/Documents/ed-pa

ck access, place your bookmarks here on the bookmarks bar. Import bookmarks now...

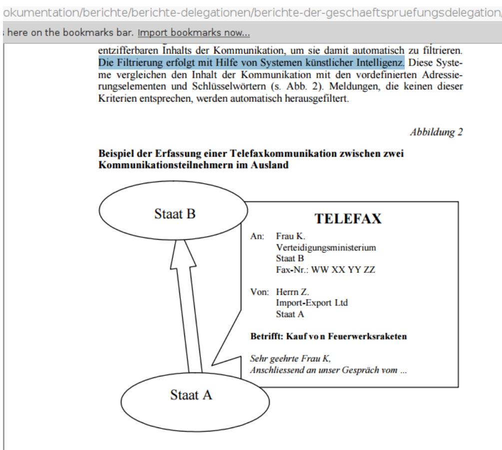
3. Februar 1995 über die Armee und die Militärverwaltung (MG; SR *510.10*), der die Aufgaben des Auslandsnachrichtendienstes im Ausland regelt.

Onyx nahm seinen Dienst im April 2000 auf und arbeitet zur Zeit im Probebetrieb. Der operationelle Betrieb wird im Laufe des Jahres 2004 aufgenommen, die Aufnahme des Vollbetriebs ist auf Ende 2005/Anfang 2006 vorgesehen.

Das Onyx-System bietet seinem hauptsächlichsten Benutzer, dem Strategischen Nachrichtendienst (SND) des Departements für Verteidigung, Bevölkerungsschutz und Sport (VBS) bereits heute zahlreiche Funktionen und Möglichkeiten der Informationsbeschaffung an. In weniger grossem Umfang dient es auch dem Dienst für Analyse und Prävention (DAP) des Eidgenössischen Justiz- und Polizeidepartements (EJPD).

Onyx ermöglicht eine Massenüberwachung von Kommunikationen. Es erleichtert die Beschaffung nutzdienlicher Informationen, beispielsweise bei der Bekämpfung der Proliferation von Massenvernichtungswaffen (Weapons of Mass Destruction [WMD]) oder des internationalen Terrorismus, wobei die diesbezüglichen Kapazitäten der Nachrichtendienste um ein Vielfaches erhöht werden.

Beispiel: Massenüberwachung Onyx (3)



okumentation/berichte/berichte-delegationen/berichte-der-geschaeftspruefungsdelegation/

here on the bookmarks bar. Import bookmarks now...

entzifferbaren Inhalts der Kommunikation, um sie damit automatisch zu filtrieren. Die Filtrierung erfolgt mit Hilfe von Systemen künstlicher Intelligenz. Diese Systeme vergleichen den Inhalt der Kommunikation mit den vordefinierten Adressierungselementen und Schlüsselwörtern (s. Abb. 2). Meldungen, die keinen dieser Kriterien entsprechen, werden automatisch herausgefiltert.

Abbildung 2

Beispiel der Erfassung einer Telefaxkommunikation zwischen zwei Kommunikationsteilnehmern im Ausland

1517

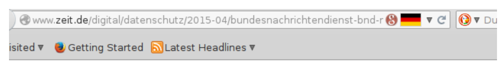
Beispiel: Massenüberwachung Onyx (4)

/.parlament.ch/d/dokumentation/berichte/berichte-delegationen/berichte-der-geschaeftspruefungsdelegation/Documents/ed-pa-gpd-o

place your bookmarks here on the bookmarks bar. Import bookmarks now...

Wenn die GPDel verlangt, dass Kommunikationsabhörungen im MG ausdrücklich festgehalten werden, verfolgt sie ein Transparenzziel. Diese Forderung rechtfertigt sich weniger auf landesinterner Ebene, da die Abhörung von Kommunikationsteilnehmern in der Schweiz verboten ist, sondern vielmehr im Hinblick auf das Völkerrecht und insbesondere auf die EMRK[41]. Artikel 8 EMRK lässt Eingriffe in die Privatleben nur dann zu, wenn es darum geht, die nationale Sicherheit zu wahren, und wenn dabei bestimmte Bedingungen wie Bestehen und Zugänglichkeit der rechtlichen Grundlage, Verhältnismässigkeit usw. erfüllt werden. Der Europäische Gerichtshof für Menschenrechte hat in mehreren Entscheiden darauf hingewiesen, dass die Gesetze zur Reglementierung administrativer oder gerichtlicher Abhörungen der Öffentlichkeit zugänglich und ausreichend genau und ausführlich abgefasst sein müssen, so dass die Bürger darauf mit einem adäquaten Verhalten reagieren können[42].

Überwachung mit Selektoren (1)



Was sind Selektoren?

Selektoren sind so etwas wie Suchbegriffe. Das können IP-Adressen, Telefonnummern, E-Mail-Adressen sein, genauso wie Geokoordinaten, MAC-Adressen, URLs. Aber auch einzelne Suchbegriffe können ein Selektor sein, also Namen oder Kürzel von Firmen und Behörden, oder Ausdrücke wie "Eurocopter".

In die Datenbanken des BND werden somit drei Dinge eingespeist: Die abgehörten Daten aus den Leitungen, die von der NSA gelieferten Selektoren und die Selektoren, die der BND selbst erstellt hat – denn auch er wühlt selbstverständlich in den Daten und sucht nach Interessantem. Als Ergebnis liefern die Rechner alle Informationen, die irgendetwas mit einem solchen Suchbegriff zu tun haben: Wen der Inhaber einer Telefonnummer angerufen hat, wer sich an einem bestimmten Ort aufgehalten hat und so weiter. Das sind die sogenannten Positiv-Selektoren, mit denen aktiv nach etwas gesucht wird.

In der Sprache des BND gibt es aber auch noch Negativ-Selektoren, die wie vorgeschaltete Filter funktionieren. Tauchen die Negativ-Selektoren in einem Suchergebnis auf, soll die weitere Analyse an dieser Stelle abgebrochen werden.

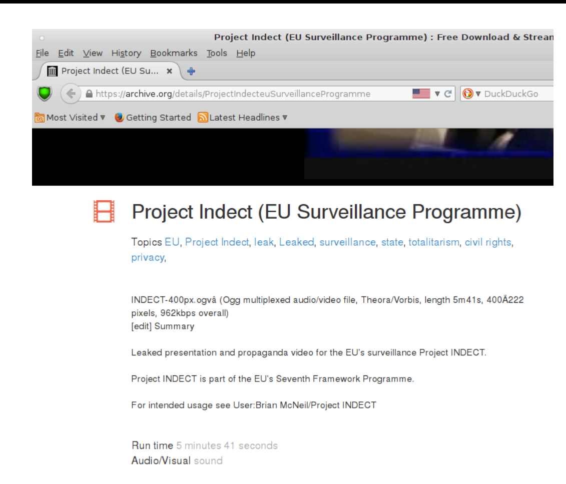
Überwachung mit Selektoren (2)

wirtschaftsblatt.at/home/nachrichten/europa/4748126/NSABND_Verdacht-von-Wirtschaftsspionage-der-USA-erhaertet

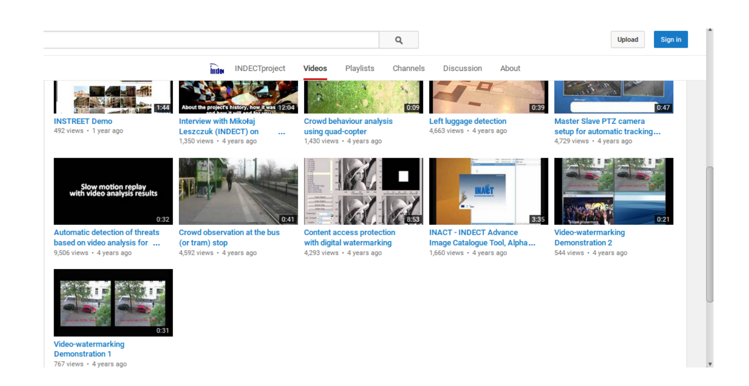
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...

Dem Bericht zufolge hatten Parlamentarier den Zeugen, der beim BND für die Prüfung und Löschung kritischer Selektoren zuständig gewesen sei, mit einer Liste von Namen aus dem Archiv des Ex-NSA-Mitarbeiters Edward Snowden konfrontiert. Unter den 31 Einträgen fänden sich Firmen wie Mercedes, Deutsche Bank, der Wertpapierdienstleister Clearstream und die Telekommunikationsfirma Debitel. Der Mitarbeiter äußerte sich den Angaben zufolge aber nicht dazu, ob und wie lange die Selektoren aktiv waren und die NSA mithilfe des BND deutsche Ziele ausgespäht hat.

Was ist INDECT? Showtime!



Mehr INDECT-Videos



Deliverable 4.1: XML Data Corpus (1)

XML Data Corpus: Report on methodology for collection, cleaning and unified representation of large textual data from various sources: news reports, weblogs, chat.

Deliverable 4.1: XML Data Corpus (2)



- **Government (Tag: ORG.GOV)**

  This subtype refers to entities that are related to governmental affairs, politics, or the state. Note that the entire government of a GPE is excluded from this subtype and should be classified as GPE.ORG as we will see later. This subtype also includes military organizations that are connected to the government of a GPE. Some examples are the following:

  *[The **British navy**] announced yesterday that . . .*

  *[The **ministry of culture**] has funded our research . . .*

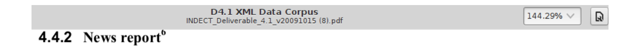
- **Commercial (Tag: ORG.COM)**

  This subtype refers to organizations, which primarily focus on providing products or services for profit. Some examples are the following:

  *[**Google's** search engine] is based on PageRank . . .*

  *[**Apple**] announced yesterday the release of its new iPhone . . .*

Deliverable 4.1: XML Data Corpus (3)



4.4.2 News report[6]

Among the most serious incidents reported to the (**National Criminal Intelligence Service** [ORG.GOV]) **NCIS** [ORG.GOV] were:

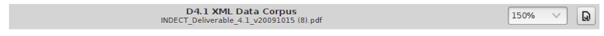
July 2008: **Glasgow Rangers** [ORG.SPO] v **Shelbourne** [ORG.SPO]. **Police baton** [ORG.GOV] charged 150 **Rangers supporters** [PER.Group] who were trying to attack fans of the **Irish club** [PER.Group].

August 2008: **Norwich City** [ORG.SPO] v QPR [ORG.SPO]: Twenty **supporters from both sides** [ORG.SPO] involved in bottle throwing in a **Norfolk** [LOC.ADD] pub. One person arrested.

30 September 2008: **Norwich City** [ORG.SPO] v **Birmingham City** [ORG.SPO]: Twenty **Birmingham fans** [PER.Group] sprayed rival supporters with **CS gas** [WEA] and attacked them with bar stools in a pub.

- **Identified Events**
  - E1: [Police baton *{charged}* 150 **Rangers supporters** who were trying to attack fans of the Irish club.]
  - E2: [150 Rangers supporters who were {*trying to attack*} **fans of the Irish club**.]
  - E3: [150 **Rangers supporters** who were {*trying to attack*} fans of the Irish club.]

Deliverable 4.1: XML Data Corpus (4)



4.4.3 Terrorist chat[7]

Shazad Tanweer [PER.Individual]: Any extra risks getting into **Pakistan** [GPE.NAT] ?

Omar Khyam [PER.Individual]: We had five **Bengalis** [GPE.NAT] last year. Guess how **we [PER.Group]** got **them** [GPE.NAT] in. From **Bangladesh** [GPE.NAT] all the way across **India** [GPE.NAT] into **Pakistan**[GPE.NAT]... **we** [PER.Group] bribed the guy [PER.Individual]. You know when you [PER.Individual] go to the check-in, it would all be set up.

Mohammed Siddique Khan [PER.Individual]: Going through the airport - normal tickets.

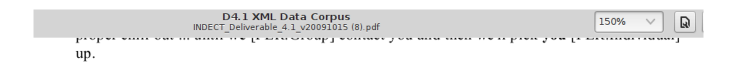
Omar Khyam[PER.Individual]: Yeah, just walk straight through bruv normal, just act as if you are a **Pakistani** [GPE.NAT].

Shazad Tanweer [PER.Individual]: I live in **Faisalbad** [GPE.NAT]

Omar Khyam [PER.Individual]: That's not a problem

Omar Khyam [PER.Individual]: All right **bruv** [PER.Individual]. Get your parents to pick you up. Or your family ... And that way you will breeze through the airport seriously. Even if **they** [ORG.GOV] are following **you** [PER.Individual] - it doesn't really count. Chill out, proper chill out ... until **we** [PER.Group] contact you and then we'll pick **you** [PER.Individual] up.

Deliverable 4.1: XML Data Corpus (5)



up.

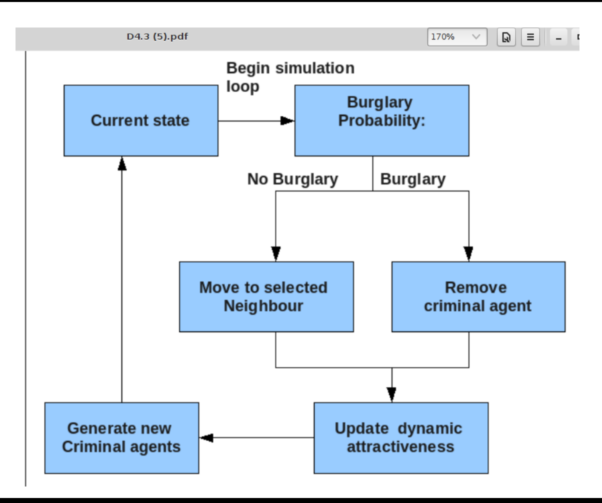
- **Identified Events**
  - E1: [Guess how **we** {got} them {in}. From Bangladesh all the way across India into Pakistan]
  - E2: [Guess how we {got} **them** {in} From Bangladesh all the way across India into Pakistan]
  - E3: **We** {bribed} the guy.
  - E4: We {bribed} the **guy**.
  - E5: [when **you** {go to the check-in}, it would all be set up.]
  - E6: [Even if **they** {are following} you]
  - E7: [Even if they {are following} **you**]

Event ID	Event Type
E1	Transportation.Illegal
E2	Transportation.Illegal
E3	FinancialTransaction.Illegal

Deliverable 4.3: Behavioural Profiling (1)

Report on current state-of-the-art on machine learning methods for behavioural profiling

Deliverable 4.3: Behavioural Profiling (2)



Deliverable 4.3: Behavioural Profiling (3)



4. Offenders characteristics profiling methods

4.1. Introduction

Modus operandi (Method of operation) is a Latin phrase that is typically used in criminal investigation to describe a crime committed by an offender as well as the methods employed for committing such a crime. The description of a crime is typically written in a police record that might contain a number of structured and unstructured data including the following:

- Free text describing the method employed for committing a crime
- Feature code for the type of a particular offence
- Feature code for the presense or absence of a specific aspect of behaviour
- Feature code for the gender of the offender
- Feature code for the age of the offender
- Feature code for the ethnic appearance of the offender

Deliverable 4.3: Behavioural Profiling (4)



4.2. Language modelling

Bache et al. [28, 2] presented a language modelling method for inferring the characteristics of offenders from an existing police archive. Based on the information included in police reports of solved cases their target was to link behavioural features with characteristics of offenders. They have focused on the following offender characteristics:

1. Gender
   The gender of a criminal agent can either be male or female.
2. Age
   The age of a criminal age is defined to be either below or above the median of the ages of offenders committing a series of crimes.
3. Ethnic appearance
   The ethnic appearance of an offender can either be white European or Afro-Caribbean.
4. Occupation
   The occupation of an offender can only take two possible values, i.e. employed or unemployed.

Deliverable 4.3: Behavioural Profiling (5)

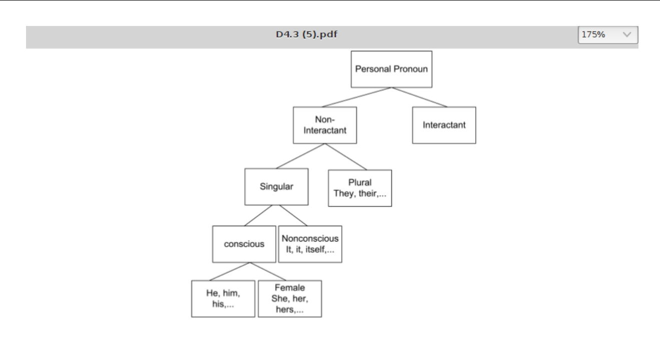


Figure 7: Functional word taxonomies

Recently, another similar approach to authorship profiling was presented in Argamon et al. [37]. Their main difference with the previous method is that their focus is not on identifying the author of a particular document. In contrast, they exploit free text in order to identify an author's gender, age, native language and neuroticism level.

Deliverable 4.11: Mine/Detect suspicious websites (1)

Specification of methods for mining and detecting suspicious websites

Deliverable 4.11: Mine/Detect suspicious websites (2)

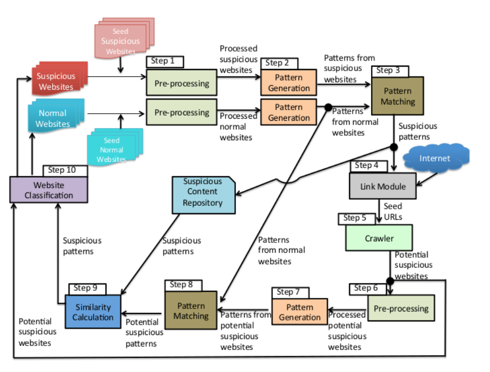
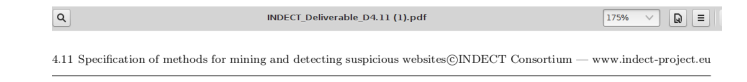


Figure 3: Method specification for detecting suspicious website

Deliverable 4.11: Mine/Detect suspicious websites (3)



N-grams

Since terms as keywords are not always accurate, phrases as keywords for document retrieval provides an alternative.. A phrase is a sequence of terms in a document. An n-gram is a continuous sequence of n terms in a text. Below we show 2-grams and 3-grams representation of the sentence, "*Hand the package to the Big Boss*".

2-gram: "*hand the*" "*the package*" "*package to*" "*to the*" "*the Big*" "*Big Boss*"

3-gram: "*hand the package*" "*the package to*" "*package to the*" "*to the Big*" "*the Big Boss*"

We can see that as we increase n we capture more context relating to words within the phrase. N-grams as features for text mining has been used to capture "semantic" information of keywords [4, 5, 6]. N-grams also have some distinctive disadvantages. Firstly, to better capture the context of any word n has to be high. Once the phrase is long, it becomes very sparse within the document collection. For example, it will be difficult to find exact match of a 6-gram phrase "*Hand the package to the Big Boss*". Also phrases contains large number of redundant terms. The terms like "*to*" will hurt the text mining process as it will be common in most type of documents.

Deliverable 4.11: Mine/Detect suspicious websites (4)

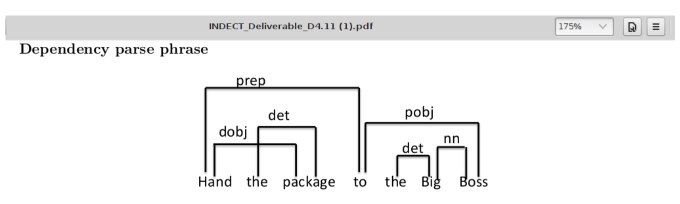
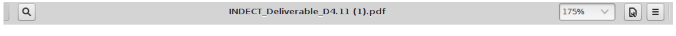


Figure 4: Dependency parse phrase for `Hand the package to the Big Boss`

A dependency parse phrase is constructed following a dependency path of given terms generated by a dependency parser. Figure 4 represents the dependency parse of the sentence **`Hand the package to the Big Boss`**. Dependency phrases have been successfully used in text classification [7] and relation mining [8]. Both text classification and relation mining share overlapped methodology with text mining. In D4.5 we successfully used dependency parse phrases for relation mining problem.

Deliverable 4.11: Mine/Detect suspicious websites (5)



6. Pattern Matching

In this step we select patterns which show high association to suspicious websites than to normal websites. In many suspicious websites, the sentences containing messages to influence criminal activities are generally grouped within other normal sentences. For example, a suspicious websites can have many factual information and few suspicious lines. Thus, the patterns extracted from such suspicious websites are not all indicative of criminal activities. Most of these patterns will also occur in normal websites. To filter out such normal patterns we use a very simple approach. Once we generate patterns from both suspicious websites and normal websites. The patterns indicative of criminal activities are only those which are not present in normal websites. Thus, we select only patterns which are present in suspicious websites but not in normal websites. For exam-

Patterns from suspicious websites	Patterns from normal website
hand-package-boss	everest-mountain
everest-mountain	tall-mountain-world
tall-mountain-world	temperature-cold-winter

Table 4: Possible patterns generated from suspicious and normal websites

Tempora

www.spiegel.de/media/media-34103.pdf

k access, place your bookmarks here on the bookmarks bar. Import bookmarks now...

[edit] A bit more detail

TEMPORA are GCHQ's large-scale, Deep Dive deployments on Special Source access (SSE). Deep Dive XKeyscores work by promoting loose categories of traffic (e.g., all web, email, social, chat, EA, VPN, VoIP...) from the bearers feeding the system and block all the high-volume, low value traffic (e.g., P2P downloads). This usually equates to ~30% of the traffic on the bearer. We keep the full sessions for 3 working days and the metadata for 30 days for you to query, using all the functionality that Keyscore offers to slice and dice the data. The aim is to put the best 7.5% of our access into TEMPORA's, comprising a mix of Deep Dive Keyscores and promotion of data based on IP subnet or technology type from across the entire MVR. At the moment, users are able to access 46x10Gs of data via existing Internet Buffers.. This is a lot of data! Not only that, but the long-running TINT program and our initial 3-month operational trial of the CPC Internet Buffer (the first operational Internet Buffer to be deployed) show that every area of ops can get real benefit from this capability, especially for target discovery and target development. Internet Buffers are different from TINT in that the latter is purely an experimental, research environment whereas Internet Buffers can be used operationally for EPR, Effects, enabling CNE etc.

For a more detailed depiction of how TEMPORA and TINT differs please see here.

XKeyscore (1)

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

What is XKEYSCORE?



1. DNI Exploitation System/Analytic Framework
2. Performs strong (e.g. email) and soft (content) selection
3. Provides real-time target activity (tipping)
4. "Rolling Buffer" of ~3 days of ALL unfiltered data seen by XKEYSCORE:
   - Stores full-take data at the collection site – indexed by meta-data
   - Provides a series of viewers for common data types

1. Federated Query system – one query scans all sites
   - Performing full-take allows analysts to find targets that were previously unknown by mining the meta-data

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

XKeyscore (2)

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

Finding Targets



- How do I find a strong-selector for a known target?
- How do I find a cell of terrorists that has no connection to known strong-selectors?
- Answer: Look for anomalous events
  - E.g. Someone whose language is out of place for the region they are in
  - Someone who is using encryption
  - Someone searching the web for suspicious stuff

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

XKeyscore (3)

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

Google Maps



- My target uses Google Maps to scope target locations – can I use this information to determine his email address? What about the web-searches – do any stand out and look suspicious?
  - XKEYSCORE extracts and databases these events including all web-based searches which can be retrospectively queried
  - No strong-selector
  - Data volume too high to forward

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

XKeyscore (4)

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

Entity Extraction



- Have technology (thanks to R6) – for English, Arabic and Chinese
- Allow queries like:
- Show me all the word documents with references to IAEO
- Show me all documents that reference Osama Bin Laden
- Will allow a 'show me more like this' capability

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

Xkeyscore (5)



https://www.tagesschau.de/inland/nsa-xkeyscore-100.html

Apps For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...

03.07.2014 18:21 Uhr

Download der Audiodatei

So schnell wird man ein "Extremist"

Siebte Stunde am Katholischen Gymnasium in Berlin-Neukölln, 13:30 Uhr: An der Wand hängen Poster von den Informatikern Tim Berners-Lee und Ada Lovelace. An die Tafel hat der Lehrer eine Zwiebel gemalt, daneben steht das Akronym "Tails".

"Tails", ist ein Betriebssystem, das das Tor-Netzwerk benutzt, um im Internet keine Spuren zu hinterlassen, das aber auch nichts vom Nutzer auf dem Computer speichert, von dem es, zum Beispiel auf einem USB-Stick, hochgefahren wird.

Darko Medic, 18, kurze braune Haare, sitzt vor seinem Laptop. Er gibt "Tails" und "USB" in die Maske seiner Suchmaschine ein. Was Darko nicht weiß: Er ist damit gerade ebenfalls in einer Datenbank der NSA gelandet. Markiert als einer der Extremisten, nach denen die Geheimdienstler so fleißig suchen.

Denn was die Regeln des Quellcodes ebenfalls verraten: Die NSA beobachtet im großen Stil die Suchanfragen weltweit - auch in Deutschland. Allein schon die einfache Suche nach Anonymisierungssoftware wie "Tails" reicht aus, um ins Raster der NSA zu geraten. Die Verbindung der Anfrage mit Suchmaschinen macht

Katalyst-Whitepaper (1)



Kapow Katalyst for OSINT

Harvest text in any language, images, audio, video from websites, blogs, and social media. Secure and non-attributable. Kapow Katalyst—*the best-kept secret in OSINT*.



Harvest Any OSINT Data with Katalyst

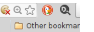
OSINT data sources are as varied as the Internet itself. Mission-critical data can reside in blogs, in news feeds, in social media—and can even be hosted on short-lived sites on the dark web. As technology standards continue to

Extract Data in Any Language

Katalyst offers built-in support for multi-byte character encodings such as Chinese and bidirectional languages such as Arabic and Hebrew. Katalyst is in daily use throughout the IC to harvest the contents of news sites,

Katalyst-Whitepaper (2)

https://www.wikileaks.org/spyfiles/docs/KAPOWSOFTWARE-2011-BuilyourOSIN-en.pdf



: access, place your bookmarks here on the bookmarks bar. Import bookmarks now... Other bookmar

Build Your OSINT Capability with Kapow

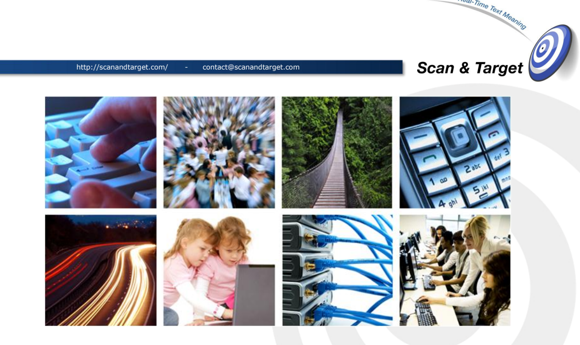
Kapow's Extraction Browser is a perfect match for the needs of OSINT.

Kapow can fully automate data extraction from any Web site. Without human intervention, Kapow Katalyst can authenticate, perform complex navigation, carry out sophisticated extraction and transformation rules, and reliably deliver data wherever it is needed.

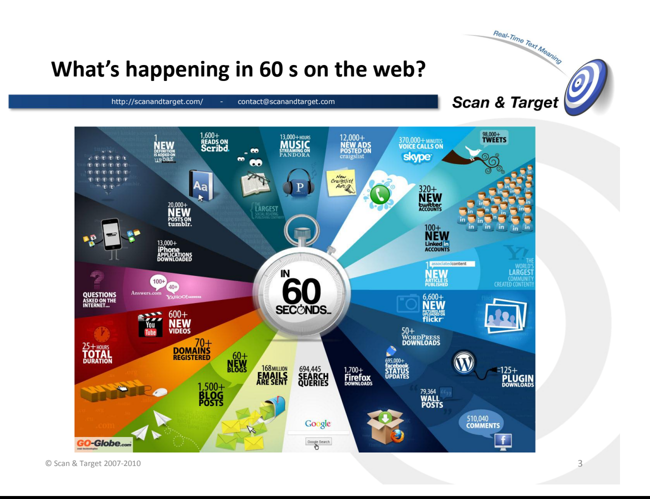
Kapow Katalyst is fully standards-based, allowing smooth exchange of data in any required form. Data can also be exchanged via Web Services, Java, or .NET functions.

Kapow's integrated environment allows you to build even the most sophisticated crawlers without the need to develop any code, and test them in real-time—increasing agility and shrinking effort to a minimum.

Scan & Target (1)



Scan & Target (2)



Scan & Target (3)

Who is Scan & Target?

http://scanandtarget.com/ - contact@scanandtarget.com



Scan & Target analyzes **digital communications in real time** to provide **actionable intelligence** to software vendors, brands, service publishers, marketing agencies, governments...



Social networks





Forums, blogs



E-mails



Instant Messaging

Our text Meaning Technology is smart enough to look in real time at an incoming text User Generated Content data stream, **see patterns of interest**, and **alert the right people** or trigger the appropriate action-- **all without being queried**

Scan & Target (4)

Scan & Target technology

http://scanandtarget.com/ - contact@scanandtarget.com



Unlike solutions based on simple keywords or semantic, our technology takes into account the **different alterations and variants** of expressions to analyze the content:

- Small/ capital letters use
- Letters repetition (vvviiiagrrra for example)
- Orthographical variations (vi@gra, vlagra, v1@gra, v149r4)
- Missing letters in some cases (v|agra, v agra…)
- Word alteration whatever the use of non alpha symbol (v.i.a.g.r.a, v_i°ag#r:a, v-iagra, viagr"a...)
- Phonetic alterations
- SMS and IM languages
- And the combination of these variations

The solution is available in **English** and **French** and **Spanish** and **Arabic (MSA + dialects, Arabic alphabet + transliteration)**.

Scan & Target (5)

Scan & Target technology

http://scanandtarget.com/ - contact@scanandtarget.com



The solution is based on a smart engine that rates not just single words but the entire content as it passes through the filtering engine. **Words** are therefore **placed in context** to **extract meaning**

The solution applies detailed thematic thesauruses - our **Smart Wordbooks**. Filters are categorized to allow customers to fine-tune the analysis (Terrorism/Drugs/Violence, etc.) according to their needs

Additional analysis layers: **sentiment analysis**, **questions detection**…

Proprietary scoring technology tailored to short digital text contents

Using a powerful and accurate conditional analysis system, our customers experience a **very low level of false positives** (between 0,05% to 0,001% in average)

Scan & Target (6)

Big Data? No problem.

http://scanandtarget.com/ - contact@scanandtarget.com



- For homeland security, our API is distributed using IBM hardware (to be hosted on your premises)
- Thanks to our connector, it's very easy to implement our API into your own applications
- You choose how to display our analysis results into your interfaces
- Capacity to deal in real time with Big Data
  - All of Twitter's traffic (10 TB / day, average 1200 Tweets per second)* could be analyzed in real time using one IBM blade center (for one language)
  - *Source - Twitter

Scan & Target (7)

Mass interception issues

http://scanandtarget.com/ - contact@scanandtarget.com



- Mass interception of digital text communications, (OSINT or COMINT like SMS, e-mails, IM…) is now technically available
- Issues for intelligence or law enforcement agencies:
  - How to deal with the volume (flow never stops)
  - How to find the needle in the digital haystack

 13

Scan & Target (8)

"Finding the needle" strategies

http://scanandtarget.com/ - contact@scanandtarget.com

Scan & Target — Real-Time Text Meaning

Benefits	Identified Suspects	Interception on keywords	Indexation and search	Text Meaning
Real time information		+	-	+
Fuzzy search	-	-		+
Advanced analysis	-	-	+	+
False positive ratio		-		+
Unknown threat detection	-		+	+
Required analyst time		-	-	+

14

Scan & Target (9)

Arabic chat alphabet

http://scanandtarget.com/ - contact@scanandtarget.com



- The Arabic chat alphabet (Arabish or Arabizi) is used to communicate in the Arabic language over the Internet or for sending messages via mobile phones when the Arabic alphabet is unavailable
- Arabic letters are replaced by letters that are phonetically equivalent
- Arabic letters that have no Latin phonetic counterpart are represented by numbers, or numbers in conjunction with an accent mark

23

Scan & Target (10)

Issues with Arabic compared to latin languages

http://scanandtarget.com/ - contact@scanandtarget.com

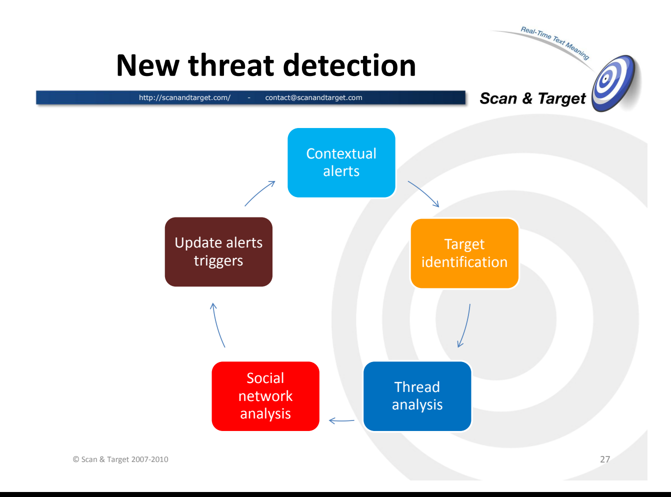


- Language identification issue:
  - MSA, dialects, mix of languages
- Transliteration issue (notably for names)
  - ABD AL-WADOUB
  - ABD EL OUADOUD
  - ABD-AL-WADUD
  - ABDEL EL-WADOUD
- Use of Arabish / Arabizi
  - bri6ania al3o'6ma / britanya al 3ozma = Great Britain for example

Our Text Meaning Technology handles all these issues

24

Scan & Target (11)



Scan & Target (12)

Messages vs thread

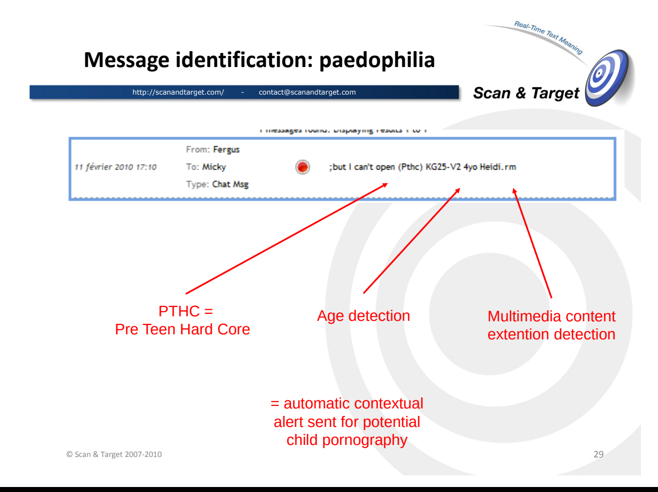
http://scanandtarget.com/ - contact@scanandtarget.com



- A web or mobile conversation is a thread of messages between 2 or more persons
- Analysis is first performed at message level for contextual alerts
- When an alert is detected, the associated discussion thread is again analyzed to:
  - Increase accuracy and precision
  - Extract investigation elements (names, places, nationality, places…)

 28

Scan & Target (13)



Scan & Target (14)

Use case: drugs traffic detection

http://scanandtarget.com/ - contact@scanandtarget.com



- Mass Surveillance of SMS communications (20 to 30 millions per day with a lot of different languages, English, Arabic, dialects…)
- Contextual alerts sent to analysts using conditional analysis on:
  - Substance related discussions,
  - Transaction related discussions (quantities, money…)
  - Middle men related discussions (dealers, luggage handler, docker, customs…)
  - Smuggling related discussions (places like ports, airports and smuggling tricks)
- Investigation by analyst (conversation thread analysis, social network analysis…) identifies:
  - Dealers' ring (pseudo, IP address…)
  - Coded language detection (use of culinary vocabulary for example)
- High precision: 40 alerts per million SMS

 31

Scan & Target (15)

Recommended solution

http://scanandtarget.com/ - contact@scanandtarget.com



- Scan & Target text meaning technology is a very efficient tool to detect previously **unknown terrorist or criminal threats** on the Internet or wireless networks
- Main benefits:
  - Ability to deal with **huge volumes** in **real time**
  - Multilingual and ability to **manage fuzzy languages like IM or arabizi**
  - **Actionable intelligence** with message & thread analysis
  - **Low level of false positive** thanks to advanced analysis
- To be integrated into your existing monitoring system

 32

Scan & Target (16)



Fragen & Antworten

